PROCEEDINGS OF TDWG

Biodiversity Information Science and Standarts 1: e20161
doi: 10.3897/tdwgproceedings.1.20161

OPEN ACCESS

Conference Abstract

Building Your Own Big Data Analysis Infrastructure for Biodiversity Science

Matthew Collins[‡,§], Nicky Nicolson[|], Jorrit Poelen[¶], Alexander Thompson[§], Jennifer Hammock[#], Anne Thessen[¤,«]

‡ University of Florida, Gainesville, United States of America
§ iDigBio, Gainesville, United States of America
| Royal Botanic Gardens, Kew, Kew, United Kingdom
¶ ManyLabs, San Francisco, United States of America
# Smithsonian Institution, Washington, United States of America
¤ The Ronin Institute for Independent Scholarship, Monclair, NJ, United States of America
« The Data Detektiv, Waltham, MA, United States of America

Corresponding author: Matthew Collins (mcollins@acis.ufl.edu)

Received: 10 Aug 2017 | Published: 10 Aug 2017

Citation: Collins M, Nicolson N, Poelen J, Thompson A, Hammock J, Thessen A (2017) Building Your Own Big Data Analysis Infrastructure for Biodiversity Science. Proceedings of TDWG 1: e20161.
https://doi.org/10.3897/tdwgproceedings.1.20161

Abstract

The size of biodiversity data sets, and the size of people's questions around them, are outgrowing the capabilities of desktop applications, single computers, and single developers. Numerous articles in the corporate sector (Delgado 2016) have been written on how much time professionals spend manipulating and formatting large data sets compared to the time they spend on the important work of doing analysis and modeling. To efficiently move large research questions forward, the biodiversity domain needs to transition towards shared infrastructure with the goal of providing a *mise en place* for researchers to do research with large data.

The GUODA (Global Unified Open Data Access) collaboration was formed to explore tools and use cases for this type of collaborative work on entire biodiversity data sets. Three key parts of that exploration have been: the software and hardware infrastructure needed to be able to work with hundreds of millions of records and terabytes of data quickly, removing the impediment of data formatting and preparation, and workflows centered around GitHub for interacting with peers in an open and collaborative manner.

We will describe our experiences building an infrastructure based on Apache Mesos, Apache Spark, HDFS, Jupyter Notebooks, Jenkins, and Github. We will also enumerate what resources are needed to do things like join millions of records, visualize patterns in whole data sets like iDigBio and the Biodiversity Heritage Library, build graph structures of billions of nodes, analyze terabytes of images, and use natural language processing to explore gigabytes of text. In addition to the hardware and software, we will describe the kinds of skills needed by staff to design, build, and use this sort of infrastructure and highlight some experiences we have with training students.

Our infrastructure is one of many that are possible. We hope that by showing the amount and type of work we have done to the wider community, other organizations can understand what they would need to speed up their research programs by developing their own collaborative computation and development environments.

Keywords

Big Data, Infrastructure, Spark, Biodiversity Informatics

Presenting author

Matthew Collins

References

- Delgado R (2016) Why Your Data Scientist Isn't Being More Inventive. Dataconomy. N.p URL: http://dataconomy.com/2016/03/why-your-datascientist-isnt-being-more-inventive/